Journal of Biomedical Informatics 44 (2011) S44–S47

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Enhancing phylogeography by improving geographical information from GenBank

Matthew Scotch [a,*], Indra Neil Sarkar [b,c,d], Changjiang Mei [a], Robert Leaman [a], Kei-Hoi Cheung [e], Pierina Ortiz [a], Ashutosh Singraur [a], Graciela Gonzalez [a]

[a] Department of Biomedical Informatics, Arizona State University, Tempe, AZ, USA
[b] Center for Clinical and Translational Science, University of Vermont, Burlington, VT, USA
[c] Department of Microbiology & Molecular Genetics, University of Vermont, Burlington, VT, USA
[d] Department of Computer Science, University of Vermont, Burlington, VT, USA
[e] Yale Center for Medical Informatics, Yale University, New Haven, CT, USA

ARTICLE INFO

Article history:
Received 14 January 2011
Accepted 13 June 2011
Available online 24 June 2011

Keywords:
Phylogeography
Databases
Nucleic acid
Geographic locations
Bioinformatics

ABSTRACT

Phylogeography is a field that focuses on the geographical lineages of species such as vertebrates or viruses. Here, geographical data, such as location of a species or viral host is as important as the sequence information extracted from the species. Together, this information can help illustrate the migration of the species over time within a geographical area, the impact of geography over the evolutionary history, or the expected population of the species within the area. Molecular sequence data from NCBI, specifically GenBank, provide an abundance of available sequence data for phylogeography. However, geographical data is inconsistently represented and sparse across GenBank entries. This can impede analysis and in situations where the geographical information is inferred, and potentially lead to erroneous results. In this paper, we describe the current state of geographical data in GenBank, and illustrate how automated processing techniques such as named entity recognition, can enhance the geographical data available for phylogeographic studies.

1. Introduction

Phylogeography is a field that focuses on the geographical lineages of species such as vertebrates or viruses [2]. Here, geographical data, such as location of a species or viral host is as important as the sequence information extracted from the species. Together, this information can help illustrate the migration of the species over time within a geographical area, the impact of geography over the evolutionary history, and the expected population of the species within the area. Molecular sequence data from NCBI, specifically GenBank [3], provide an abundance of available sequence data for phylogeographical studies. However, geographical data is inconsistently represented and sparse across GenBank entries. This can impede analysis and in certain situations lead to potentially erroneous results. In the latter, an example could be an epidemiologist attempting to track Leptospirosis in rats. The results of the phylogeographical study suggest that a surge in host infection has just occurred. In the case where only state-level geography is provided, the epidemiologist might conclude that this is a statewide issue that requires a lot of money. In actuality, had the real locations been known, the results would suggest vicariance and limitation of spread to only one small area of the state. Hawaii is an example where this could occur. This is an extreme case, but highlights the fact that lack of sufficient geographic data may distort results.

While there have been different approaches to integrating geography within evolutionary models, a recent approach by Lemey et al. [10] using the BEAST program [5] has gained much attention. BEAST uses Bayesian inferences using Markov chain Monte Carlo (MCMC) to estimate dispersal throughout the entire evolutionary model, with geographical locations represented as discrete states (like residues in a sequence) [10]. The observed geographical data is represented in the tips of the Bayesian phylogenetic tree, and the MCMC method along with a substitution matrix is used to estimate the ancestral states and the migration paths along the branches [10]. In another example, Wallace and Fitch [20] studied the phylogeography of H1N1 in various animal hosts examining migration over Europe, Asia, and Africa. The authors used GenBank records and records from the Influenza Sequence Database [12] that included many different viral hosts from different locales [20]. Like Lemey et al. [10], geographic information was assigned to the tips of a constructed tree by representing the information as multiple states [20]. PAUP* [18] was used to construct a maximum

* Corresponding author. Address: Arizona State University, 425 N. 5th St., ABC Building, Phoenix, AZ 85004, USA. Fax: +1 602 827 2564.
E-mail addresses: matthew.scotch@asu.edu (M. Scotch), neil.sarkar@uvm.edu (I.N. Sarkar), changjiang.mei@asu.edu (C. Mei), bob.leaman@asu.edu (R. Leaman), kei.cheung@yale.edu (K.-H. Cheung), ashutosh.singraur@asu.edu (A. Singraur), graciela.gonzalez@asu.edu (G. Gonzalez).

1532-0464
doi:10.1016/j.jbi.2011.06.005

parsimony tree that assigned geographical locations to the ancestral nodes [20] and MigraPhyla [21] to calculate the number of migration events among the locations [20].

These studies relied on geographical information in their analysis and phylogeography. Unfortunately, the amount of geographical information in GenBank is underreported in relation to the amount of data that actually exist for a given molecular sequence. This makes phylogeography difficult since the researcher is forced to review the paper (or other source document[s]) for any geographical information and can then link this data to the individual isolates. This can be time consuming and laborious, not to mention that it is often not possible depending on how much information is presented in the text and the supplementary tables.

One potential solution for enhancing geographical data for phylogeography is to access the journal article of each molecular study being analyzed. Then using statistical techniques, geographic and genomic information can be linked to the appropriate GenBank record. Named entity recognition (NER) might support this process. NER is a low-level information extraction method consisting of locating entities such as proteins, cell lines or diseases within natural language text and specifying their semantic type. Systems for performing named entity recognition are typically used as part of a larger pipeline for text mining tasks, such as extracting protein–protein interactions. These systems typically use lexical match, a manually created rule set, machine learning, or a combination. For example BANNER [9], and ABNER [16] perform name entity recognition of gene names in biomedical text.

In this paper, we describe the current state of geographical data for a subset of GenBank records and then highlight how named entity recognition methods can enhance the amount of geographical data available for phylogeography research. For this work, we focus on phylogeography of viruses [7], an emerging area of concentration in the field. Specifically we consider RNA viruses within tetrapod hosts as these represent the variables for many zoonoses; infectious diseases transmittable between animals and humans [8].

2. Materials and methods

We first estimated the number of sequences in GenBank that did not contain a sufficient level of geography. For the purposes of this work, we defined *a priori*, the sufficient level to be anything below state/provincial level. For example, we considered a sequence record with *New Hampshire* to be insufficient, but a sequence record with *Bedford, NH* (a town) to be sufficient. This cutoff was somewhat arbitrary, but based on the need to get more of an accurate assessment of virus location within a state (especially with larger states such as Texas, California, and Alaska). In addition, we focused on the phylogeography of RNA viruses within animal hosts, thus we only considered these GenBank records.

One of the authors (INS), who maintains a local database of NCBI GenBank metadata, extracted the records using the following process:

1. Retrieve all the sequences from GenBank that are either single-stranded RNA (ssRNA) viruses or double-stranded RNA (dsRNA) viruses (NCBI Taxonomy IDs 439488 and 35325, respectively).
2. Filter results from #1 down to just those that had an identifiable host in the host field and the host could be mapped to a tetrapod (i.e., NCBI Taxonomy ID 32523) in the lineage.
3. Split out the results into four categories based on the contents of the '*country*' and '*lat_lon*' fields (having either or both fields populated).

In GenBank, the *country* field varies in terms of granularity. For example, some entries have just a country (e.g. USA), while other entries are much more detailed (e.g. Japan:Saitama, Tokorozawa). The *lat_lon* field indicates the latitude and longitude coordinate for the location. A GenBank record can include a *lat_lon* field with or without the *country* field.

In order to identify which records had a sufficient level of geography, one of the authors (CM) linked the records to a local database that contained information from GeoNames.org [19], an online resource of geographical entities. States and provinces in GeoNames are 1st-level administrative boundaries (ADM1), while counties (ADM2), municipalities, and towns have a different administrative code. For this study, we considered any geographical boundary below ADM1 (except for regions) to be *sufficient*.

After we assigned the extracted GenBank records as *sufficient*/*insufficient*, two judges manually scored a random sample of 500, and this served as the gold standard to evaluate the categorization.

In addition, we randomly selected 10 GenBank records deemed to have insufficient geography detail and downloaded their full-text articles from PubMed [13]. One of the authors (MS) annotated them using Knowtator [14]. We included tables, supplementary information, and any captions or labels from tables or figures (but not the figures themselves). The annotator looked for geographical entities (towns, states, countries, etc.) in each of the 10 papers, as well as gene names. The purpose of this effort is to eventually link the genomic information to a geographical location, and then back to the GenBank record.

We used two NER systems, BANNER for genes, and the Stanford Named Entity Recognizer system [6] for geographic NER, on the annotated documents. BANNER is based on conditional random fields, a type of probabilistic modeling. It was designed for extraction of gene names from journal abstracts [9]. We chose BANNER because of its success in information extraction in bioinformatics [9] as well as technical knowledge by two of the developers (RL and GG). The Stanford NER is also based on conditional random fields and written in Java. We used it because of familiarity by the authors and our belief that it could reasonably handle NER of geography names from this domain. We used the 3-class trained model distributed with the Stanford NER code. We trained this model on a combination of several different US and UK newswire corpora, and labels PERSON, ORGANIZATION and LOCATION entities, though we only utilized LOCATION. We also applied three filters on BANNER in order to improve precision. The filters included: dropping any mention that starts with "fig" since BANNER interpreted many figure captions as gene names, dropping any mention with multiple adjacent spaces, since these are typically multiple cells from a table in the original document, dropping any mention longer than 20 characters, for example "HLA-Associated Viral Polymorphisms".

We formatted the documents using Adobe Acrobat [1] to transform them from pdf to plain text. We also performed manual cleaning of the documents such as sentence splitting and removal of unusual characters as a result of converting from pdf to text.

3. Results

The search for dsRNA or ssRNA viruses within a tetrapod host returned 105,652 GenBank records (Table 1). We found 21,003 of the 105,652 (20%) to be sufficient based on our definition, while 84,601 (77,865 + 7736) (80%) were insufficient.

We linked the records using SQL to the GeoNames data in another local database. We performed manual inspection when there were issues in linking the records. For example, less than 1% of the records did not exist in our local database, requiring manual inspection and 48 could not be assigned (<1%). In order to evaluate this entire approach, two of the authors (PO, AS) reviewed a random sample of 500 records, reached a consensus on any disagree-

Table 1
Analysis of geography data in GenBank. Search for ssRNA viruses or dsRNA viruses (NCBI Taxonomy IDs 439488 and 35325) in a tetrapod host (NCBI Taxonomy IDs 32523).

Numerator	Count	%
No geography data (insufficient)	7736	7
Geography data ⩾ ADM1 OR a region (insufficient)	76,865	73
Geography data < ADM1 OR not a region (sufficient)	21,003	20
Unlinked records	48	<1
Total	105,652	

Table 2
Evaluation of NER systems for geography and gene names.

Measure	Stanford (geography)	BANNER (genes)
Precision	0.81	0.40
Recall	0.26	0.45
F-measure | 0.39 | 0.42 |

ments, and produced a gold standard set for comparison. This entire process took multiple iterations in order to address errors that arose. For example, we discovered after the first iteration that our data contained records of non-tetrapod hosts. We modified the search filter to correct this error and re-classified all of the records in the new dataset. For our second iteration, comparison of the two reviewers' gold standard set to the data set showed poor sensitivity and specificity (.40 and .20 respectively). In light of these results, we reviewed the linking with the Geonames data and discovered several misclassifications. For example, we categorized many abbreviated locations such as 'USA:SC' as populated places (PPL) rather than ADM1. This is mostly due to the inconsistencies with the location information in GenBank. In this instance, our linkage retrieved 'PPL' and classified the location (South Carolina) as sufficient. We attempted a third iteration to correct these issues and recalculated the final results that are shown in Table 1.

The final inter-rater reliability was 0.86 and calculated using Cohen's κ [4]. The final sensitivity of categorizing the GenBank records as sufficient/insufficient was 0.97 and the specificity was 0.70.

Table 2 shows the results of the two NER systems for extraction of geography terms and gene names. The table shows the results for the third and final iteration. We also performed NER for the initial iteration, but we discarded these results since some of the 10 articles did not meet our inclusion criteria. In addition, researchers had trained BANNER and the Stanford NER on data sets during prior research work and thus we did not do any additional training during this round.

For geography NER, the Stanford tool used strict matching, meaning that the system must find both the correct left boundary and right boundary for it to be considered correct. The low numbers for both systems suggest poor recognition of both geography and gene entities from these papers.

4. Discussion

Both the Stanford NER and BANNER performed poorly in their tasks over the 10 papers. Annotation errors certainly contributed to their performance as only one annotator was used during this round. However, additional explanations vary for each system.

The Stanford NER system had reasonable precision but poor recall. We analyzed the errors and found three main issues. The first is that the articles contained many place names in tables or a long list of abbreviations. As the Stanford system was trained on text in full sentences, performance suffered on names in other contexts. The second is that the authors often introduced their own abbreviation for a place name, such as "HLJ" for "Heilongjiang." While there has been some prior work on using such local abbreviations [15], most of these assume that both the abbreviation and its expanded form will appear in a specific format that uses parenthesis to mark one of the forms, such as "Heilongjiang (HLJ)." We did not encounter this form in the articles annotated for this study. The third reason is that many of the places in the articles annotated are cities or towns rather than nations or continents. Since there are so many more cities than nations, the Stanford NER system would not have seen the majority of them previously and would therefore be required to rely on context, missing many of them.

For BANNER, the poor performance is primarily due to the differences between the dataset used to train the model for BANNER and the articles used in this study. The model used by BANNER was learned from the training data for the BioCreative 2 Gene Mention task [17], which consists of 15,000 sentences from PubMed abstracts. We found that the articles for the present study contained far more abbreviated names for genes and proteins than the full names which are often used in PubMed abstracts, causing BANNER to miss many of these. Second, the annotation criteria for the BioCreative 2 Gene Mention data included a specificity requirement so that if the annotator could not determine which specific gene or protein was being mentioned it was not annotated. BANNER would therefore not tag mentions, such as "human leukocyte antigens," which could refer to more than one gene or protein. In addition, the articles for this study contain many short names in ambiguous contexts that appear to be morphologically similar to gene and protein names.

In addition to our NER evaluation, we also described the geography data included in GenBank and found a high proportion of insufficient detail. One reason for this might be that many researchers do not consider this data an important element in their research. The result is that they either did not specify it at all during submission to GenBank or provide only a 'vague' description (e.g. USA). In addition to impeding research, the results from phylogeographic studies that use insufficient geography may be misleading and cause unfortunate consequences. These include inappropriate public health action including the spending of valuable tax dollars, or, lack of public health action and an increase in the number infected with the virus.

In the field of biomedical informatics, the use of named entity recognition offers a solution to this problem by extracting geographical and genomic data from the free text and tables of the journal articles. Once the data has been extracted, statistical methods have the potential to link a GenBank entry to a more precise geographical location. This would provide the researcher with robust dataset for their phylogeographical analysis. In addition, natural language processing and NER can support ontology enrichment [11], allowing for phylogeography-related data to be used for other research projects.

5. Limitations

There are several limitations to this work. The assumption that 1st-level administrative boundaries (a state/province) are an insufficient level of geography might not always be the case. For example, small states or provinces might find this information to be sufficient. Ultimately the choice of what geographical level is sufficient is dependent on the research question being pursued.

Another limitation is the low proportion of GenBank records that had PubMed links. For example, using an online search filter through NCBI, we estimated that 42% of GenBank sequences in our result set of 105,652 had PubMed records. Using the same filter, we estimated that 46% of the dsRNA and ssRNA records online

as of April 4, 2011 had PubMed records. This could result in bias since less than half of the records had links to PubMed, and even less had links to full-text online articles.

A final limitation is that only one annotator was used for the NER work. As described, annotation errors likely contributed to the poor performance of both NER systems. Ideally more than one annotator should be used and inter-rater agreement assessed. Issues with time and identifying a second annotator prevented this. Also the small amount of annotated papers, 10, was mainly due to issues with time and resources.

6. Conclusion

The field of phylogeography requires combining geographical information related to locations of species, with the evolutionary history of those species. NCBI databases such as GenBank are one of the main portals for accessing genomic data for use in development of evolutionary models. Unfortunately, our analysis indicated a lack of sufficient geographical data in GenBank. Named entity recognition provides an opportunity to enhance the amount of geographical data available and the expansion of genomic data available for phylogeography.

Additional work will focus on training the NERs for improving the performance of genes and geography recognition, as well as incorporating additional classes such as strains, isolates, and accession numbers. Once this has occurred, statistical mapping will be done of genes, accession numbers, isolates, and strains to the geographical information found in the paper. Finally, we plan to develop a Web service to automate this process and easily provide the researcher with this enriched data.

Acknowledgments

This research was supported in part by National Institutes of Health/National Library of Medicine (NLM) Grants R00LM009825 (to MS) and R01LM009725 (to INS).

References

[1] Adobe Acrobat; 2010.
[2] Avise JC. Phylogeography: the history and formation of species. Cambridge (MA): Harvard University Press; 2000.
[3] Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucl Acids Res 2008;36:D25–30.
[4] Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Measure 1960;20:37–46.
[5] Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 2007;7:214.
[6] Finkel JR, Grenager T, Manning C. Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL 2005); 2005. p. 363–70.
[7] Holmes EC. The phylogeography of human viruses. Mol Ecol 2004;13:745–56.
[8] Krauss H. Zoonoses: infectious diseases transmissible from animals to humans. 3rd ed. Washington (DC): ASM Press; 2003.
[9] Leaman R, Gonzalez PL. BANNER: an executable survey of advances in biomedical named entity recognition. Pacific Symp Biocomput 2008:652–63.
[10] Lemey P, Rambaut A, Drummond AJ, Suchard MA. Bayesian phylogeography finds its roots. PLoS Comput Biol 2009;5:e1000520.
[11] Liu K, Hogan WR, Crowley RS. Natural language processing methods and systems for biomedical ontology learning. J Biomed Inform 2011;44:163–79.
[12] Macken C, Lu H, Goodman J, Boykin L. The value of a database in surveillance and vaccine selection. Amsterdam: Elsevier; 2001.
[13] NCBI, PubMed; 2011.
[14] Ogren PV. Knowtator: a protégé plug-in for annotated corpus construction. In: Proceedings of the 2006 conference of the north American chapter of the association for computational linguistics on human language technology; 2006. p. 273–5.
[15] Schwartz AS, Hearst MA. A simple algorithm for identifying abbreviation definitions in biomedical text. Pacific Symp Biocomput 2003:451–62.
[16] Settles B. Biomedical named entity recognition using conditional random fields and rich feature sets. In: Proceedings of the COLING international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP); 2004.
[17] Smith L, Tanabe LK, Ando RJ, Kuo CJ, Chung IF, Hsu CN, et al. Overview of BioCreative II gene mention recognition. Genome Biol 2008;9(Suppl. 2):S2.
[18] Swofford D. PAUP*: phylogenetic analysis using parsimony (and other methods) 4.0 beta for Windows/DOS, Sinauer; 2002.
[19] Vatant B, Wick M; 2007.
[20] Wallace RG, Fitch WM. Influenza A H5N1 immigration is filtered out at some international borders. PLoS One 2008;3:e1697.
[21] Wallace RG, Hodac H, Lathrop RH, Fitch WM. A statistical phylogeography of influenza A H5N1. Proc Natl Acad Sci USA 2007;104:4473–8.